MuMod: A Micro-Unit Connection Approach for Hybrid-Order Community Detection

Ling Huang,[1,2] **Hong-Yang Chao,**[1,2] **Guangqiang Xie**[3]
[1]School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
[2]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
[3]School of Computer, Guangdong University of Technology, Guangzhou, China
huanglinghl@hotmail.com, isschhy@mail.sysu.edu.cn, xiegq@gdut.edu.cn

Abstract

In the past few years, higher-order community detection has drawn an increasing amount of attention. Compared with the lower-order approaches that rely on the connectivity pattern of individual nodes and edges, the higher-order approaches discover communities by leveraging the higher-order connectivity pattern via constructing a motif-based hypergraph. Despite success in capturing the building blocks of complex networks, recent study has shown that the higher-order approaches unavoidably suffer from the hypergraph fragmentation issue. Although an edge enhancement strategy has been designed previously to address this issue, adding additional edges may corrupt the original lower-order connectivity pattern. To this end, this paper defines a new problem of community detection, namely hybrid-order community detection, which aims to discover communities by simultaneously leveraging the lower-order connectivity pattern and the higher-order connectivity pattern. For addressing this new problem, a new Micro-unit Modularity (MuMod) approach is designed. The basic idea lies in constructing a micro-unit connection network, where both of the lower-order connectivity pattern and the higher-order connectivity pattern are utilized. And then a new micro-unit modularity model is proposed for generating the micro-unit groups, from which the overlapping community structure of the original network can be derived. Extensive experiments are conducted on five real-world networks. Comparison results with twelve existing approaches confirm the effectiveness of the proposed method.

Introduction

Community detection is a hot research topic in network mining. Many community detection approaches have been developed from different perspectives (Wang, Lai, and Yu 2013; 2014; Shao et al. 2015; He et al. 2016; Blondel et al. 2008; He et al. 2017; 2018; Jin et al. 2018; Sun et al. 2018; Li et al. 2018a; Ganji, Bailey, and Stuckey 2018; Jin et al. 2019; Laishram, Wendt, and Soundarajan 2019; Wang and Zhu 2019; Zhang et al. 2019). According to the adopted connectivity pattern, the existing community detection approaches can be roughly categorized into two classes, namely lower-order community detection and higher-order community detection. In the lower-order community detection approaches, only the lower-order connectivity pattern is utilized (Shi and Malik 2000; Newman 2006; Frey and Dueck 2007; Schaeffer 2007; Chakraborty et al. 2014), which can be captured at the level of individual nodes and edges. By ignoring the higher-order connectivity pattern, the community structure discovered by the lower-order approaches fails to capture the building blocks of complex network (Benson, Gleich, and Leskovec 2016).

On the other hand, the existing higher-order community detection approaches mainly rely on the higher-order connectivity pattern at the level of small subnetworks (Arenas et al. 2008; Zhao 2015; Benson, Gleich, and Leskovec 2016; Tsourakakis, Pachocki, and Mitzenmacher 2017; Zhou et al. 2017; Yin et al. 2017; Huang, Wang, and Chao 2018a; Li et al. 2019a; 2018b; Huang, Wang, and Chao 2019). In the higher-order approaches, a motif-based hypergraph is constructed by utilizing only the co-occurrence information of the motif instances. However, as shown in (Li et al. 2019b), it may encounter the hypergraph fragmentation issue, in which the original connected network may be fragmented into a large number of connected components with various sizes and isolated nodes due to the lack of higher-order connection among them. Although an edge enhancement approach has been developed for addressing this issue (Li et al. 2019b), our study shows that adding additional edges may corrupt the original lower-order connectivity pattern. To our best knowledge, there is still a lack of approaches designed for effectively leveraging both of the lower-order connectivity pattern and the higher-order connectivity pattern.

In this paper, a new problem of community detection is defined, namely hybrid-order community detection, which aims to discover communities by simultaneously leveraging the lower-order connectivity pattern and the higher-order connectivity pattern. For addressing this new problem, a new Micro-unit Modularity (MuMod) approach is designed. Firstly, a new concept called micro-unit is defined, which is either a motif instance or an edge that is not contained in any motif instance. The micro-unit connection of two micro-units is defined to be the Jaccard similarity between their node sets. Then, a micro-unit connection network is constructed, where the micro-units are regarded as nodes and the micro-unit connections are regarded as weighted edges.

Finally, a new micro-unit modularity model is proposed for generating the micro-unit groups, from which the overlapping community structure of the original network can be derived. Extensive experiments are conducted on five real-world networks. Comparison results with twelve existing approaches confirm the effectiveness of the proposed method.

Background and Problem Statement

The input is an undirected and unweighted network $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ consisting of n nodes $\mathcal{V} = \{v_1, \ldots, v_n\}$ and m edges $\mathcal{E} = \{e_1, \ldots, e_m\}$. In this work, we focus on the scenario that the edges are undirected and unweighted. However, our technique can be easily extended to other scenarios as well. Before formally defining the problem of hybrid-order community detection, we first briefly introduce the background and some notations.

Different from the lower-order connectivity pattern that can be captured at the level of individual nodes and edges, i.e. an edge e_l connecting two nodes v_i and v_j, the higher-order connectivity pattern is a more complex structure, i.e. a subnetwork consisting of more than one edges and the corresponding ending nodes. It is regarded as the building blocks of complex network (Benson, Gleich, and Leskovec 2016).

One representative higher-order structure is motif, which is defined as follows (Milo et al. 2002).

Definition 1 (Motif) *Motif is a dense subnetwork occurring in complex networks at numbers that are significantly higher than those in randomized networks preserving the same degree of nodes. It is denoted as* $\mathbf{M} = \{\mathcal{V}_\mathbf{M}, \mathcal{E}_\mathbf{M}\}$ *where* $\mathcal{V}_\mathbf{M}$ *and* $\mathcal{E}_\mathbf{M}$ *denote the node set consisting of* p *nodes and edge set consisting of* q *edges in the motif* $\mathbf{M}$ *respectively, with* q *being between* $p-1$ *(a tree motif) and* $\frac{p(p-1)}{2}$ *(a clique motif).*

Usually, the following Z-score is adopted to identify the statistically significant motifs in one network (Milo et al. 2004),

$$Z = \frac{N_{real} - \text{mean}(N_{rand})}{\text{std}(N_{rand})} \tag{1}$$

where N_{real} is the number of occurrences of the subnetwork in the real network, $\text{mean}(N_{rand})$ and $\text{std}(N_{rand})$ represent the mean and standard deviation of the numbers of occurrences of the subnetwork in the r randomly rewired networks preserving the same node degrees. In our experiment, following the conventional setting in the literature (Lin et al. 2017), the number of randomly rewired networks, i.e. r, is set to be 1000.

In the higher-order community detection approaches, usually, only the subnetwork with the highest Z-score is identified as the motif and utilized. Obviously, different motifs may be discovered from various types of networks (Milo et al. 2002). We can even define one "consensus motif" as the motifs shared by one type of networks. The most widely discovered and studied motifs are 3-node triangle motif and 4-node motif consisting of 3 nodes and 4 nodes respectively, due to their wide appearance in diverse networks as building blocks. In this work, following the conventional setting (Tsourakakis, Pachocki, and Mitzenmacher 2017; Li et al. 2019b), we focus on the 3-node triangle motif, but our technique can be easily extended to other motifs as well.

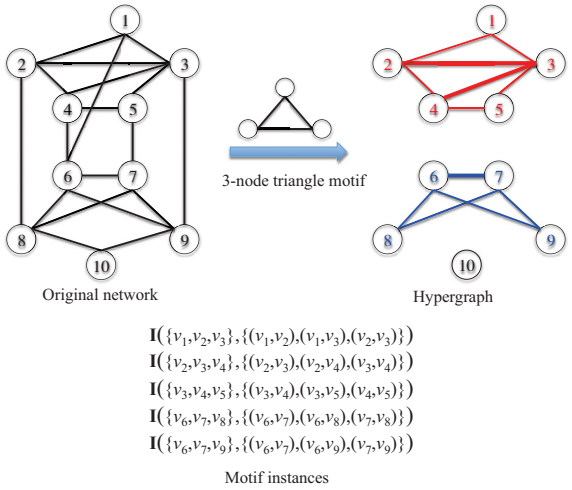


Figure 1: Illustration of the 3-node triangle motif and the corresponding hypergraph. For clarity, the five motif instances are also listed. The hypergraph is fragmented into two connected components and one isolated node, as plotted in red, blue and black respectively.

After identifying the motif, all the motif instances are searched from the network to form a motif instance set, which is formally defined as follows.

Definition 2 (Motif Instance) *The motif instance* $\mathbf{I}(\{v_{x_1}, \ldots, v_{x_p}\}, \{e_{y_1}, \ldots, e_{y_q}\})$ *is a specific appearance of motif* $\mathbf{M}$ *in the network. That is,* $|\{v_{x_1}, \ldots, v_{x_p}\}| = |\mathcal{V}_\mathbf{M}|$ *and the edge set* $\{e_{y_1}, \ldots, e_{y_q}\}$ *has the same topological structure as* $\mathcal{E}_\mathbf{M}$.

Definition 3 (Motif Instance Set) *The motif instance set* $\mathcal{M} = \{\mathbf{I}_1, \ldots, \mathbf{I}_{\bar{m}}\}$ *consists of all the motif instances appearing in the network. That is,* $\mathbf{I}_i, \forall i = 1, \ldots, \bar{m}$ *is a motif instance defined in Definition 2.*

Figure 1 plots the 3-node triangle motif instance set found in a network consisting of 10 nodes. In this example, the five motif instances construct a hypergraph, in which the nodes are the same as the original network but the edge represents the number of motif instances simultaneously containing the two ending nodes. From the figure, we can see that, the hypergraph is fragmented into two connected components and one isolated node due to the reason that only the motif instances are utilized without considering the lower-order connectivity pattern (i.e. edges). Although this simple illustration looks like that the fragmentation of hypergraph is a good partition of the original network, it is not true in larger networks. As illustrated in (Li et al. 2019b), for a real-world connected network, the hypergraph is often fragmented into several connected components with various sizes and a large number of isolated nodes. Obviously, it is not suitable to directly use such fragmented hypergraph for higher-order community detection, since it would result in the over-partitioning of the original network. Unfortunately, such hypergraph fragmentation issue is often ignored

in the existing higher-order approaches. The main reason for the hypergraph fragmentation issue is that, only the higher-order connectivity pattern is utilized, which may be much sparser than the original lower-order connectivity pattern (i.e. edges). Therefore, some fragmentations would appear in the case of missing higher-order connection.

Although an edge enhancement approach was developed for addressing this issue (Li et al. 2019b), adding additional edges may corrupt the original lower-order connectivity pattern. To our best knowledge, there is still a lack of approaches designed for effectively leveraging both of the lower-order connectivity pattern and the higher-order connectivity pattern. To this end, in this paper, we propose a novel community detection problem called hybrid-order community detection, which is formally defined as follows.

Definition 4 (Hybrid-order Community Detection) *Hybrid-order community detection aims to discover community structure by leveraging not only the lower-order connectivity pattern but also the higher-order connectivity pattern.*

The Proposed Approach

In this section, we will describe the Micro-unit Modularity (MuMod) approach for hybrid-order community detection.

Micro-unit Connection Network Construction

To simultaneously leverage the lower-order connectivity pattern and the higher-order connectivity pattern, both of the motif instances and the edges should be considered during the community detection. A naive approach is to directly design a weighted similarity matrix with each entry representing the sum of the corresponding entries from the original adjacency matrix and the hypergraph similarity matrix. And then, based on the weighted similarity matrix, some existing similarity based community detection approaches can be applied, such as spectral clustering (Shi and Malik 2000) and modularity (Newman 2006). However, directly integrating the original adjacency matrix and the hypergraph similarity matrix would result in the replication of some edges, which, like the edge enhancement strategy (Li et al. 2019b), may corrupt the underlying community structure.

To this end, we propose a novel concept called micro-unit, which effectively encodes both of the lower-order connectivity pattern and the higher-order connectivity pattern.

Definition 5 (Micro-unit) *A micro-unit u is defined as either a motif instance or an edge that is not contained in any motif instance. And the node set and the edge set of a micro-unit u are denoted as $\mathcal{V}^u$ and $\mathcal{E}^u$ respectively.*

1. If a micro-unit u is a motif instance, then $u = \mathbf{I}(\{v_{x_1}, \ldots, v_{x_p}\}, \{e_{y_1}, \ldots, e_{y_q}\})$ with $\mathcal{V}^u = \{v_{x_1}, \ldots, v_{x_p}\}$ and $\mathcal{E}^u = \{e_{y_1}, \ldots, e_{y_q}\}$ respectively.
2. If a micro-unit u is an edge, then $u = \{\{v_{i'}, v_{i''}\}, \{e_i\}\}$ with $\mathcal{V}^u = \{v_{i'}, v_{i''}\}$ and $\mathcal{E}^u = \{e_i\}$, where $v_{i'}$ and $v_{i''}$ are the two ending nodes of e_i in the original network.

Based on the concept of micro-unit, the micro-unit connection can be formally defined as follows.

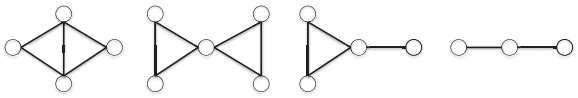


Figure 2: Illustration of all the possible types of micro-unit connections in the case of 3-node triangle motif. Notice that in Case I and Case II, although the two micro-units are motif instances, different connecting manners lead to different connection strengths.

Definition 6 (Micro-unit Connection) *Two micro-units are connected if they share at least one common nodes. Let u_i and u_j denote two micro-units, their connection strength is defined as the Jaccard similarity of their node sets*

$$Sim(u_i, u_j) = \frac{|\mathcal{V}^{u_i} \bigcap \mathcal{V}^{u_j}|}{|\mathcal{V}^{u_i} \bigcup \mathcal{V}^{u_j}|}. \tag{2}$$

Notice that, although the Jaccard similarity is adopted for measuring the connection strength in Definition 6, other similarity measures can be adopted for measuring the connection strength of two micro-units. For instance, in the case of non-isomorphic motifs (e.g. a 4-node motif consisting of 5 edges), different nodes should be emphasized differently. Nevertheless, our technique can be easily extended to other scenarios.

As a simple example, Figure 2 illustrates all the possible types of micro-unit connections in the case of 3-node triangle motif. From the figure, it is clear that, in the case of 3-node triangle motif, the micro-unit connection strengths take values only from the set $\{\frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}\}$.

Based on the micro-units and micro-unit connections, a micro-unit connection network can be constructed, where the micro-units are regarded as nodes and the micro-unit connections are regarded as weighted edges.

Definition 7 (Micro-unit Connection Network) *A micro-unit connection network is a network where the micro-units are regarded as nodes and the micro-unit connections are regarded as edges. In particular, the topological structure is represented by the micro-unit connection matrix $A \in \mathbb{R}^{|\{u\}| \times |\{u\}|}$, where $|\{u\}|$ denotes the number of micro-units. Each A_{ij} denotes the micro-unit connection of micro-units u_i and u_j, $i \neq j$, i.e.*

$$A_{ij} = \begin{cases} Sim(u_i, u_j), & \text{if } u_i \text{ and } u_j \text{ are connected} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Hereafter, we will use A to denote the micro-unit connection network. For illustration purpose, Figure 3 plots all the micro-units and the micro-unit connection matrix obtained from the original network shown in Figure 1. It is obvious that both of the lower-order connectivity pattern and the higher-order connectivity pattern of the original network are encoded in the micro-unit connection network. Compared with the hypergraph constructed by only the higher-order connectivity pattern, the micro-unit connection network is able to overcome the hypergraph fragmentation issue.

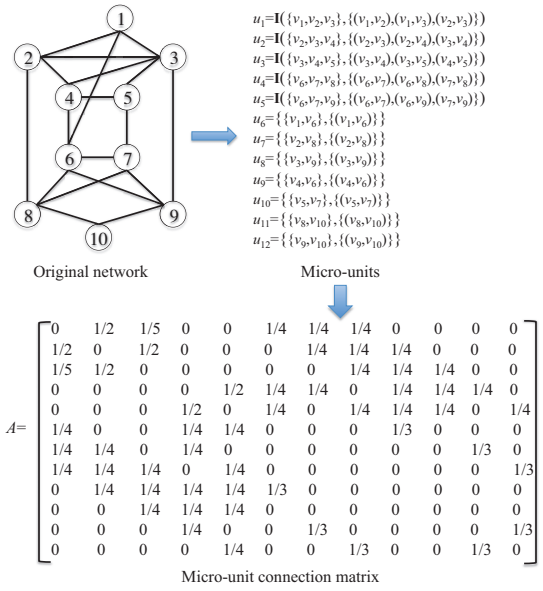


Figure 3: Illustration of all the micro-units and the micro-unit connection matrix obtained from the original network shown in Figure 1.

Notice that different from the original network, the nodes in the micro-unit connection network are micro-units rather than the original singleton nodes. Based on the micro-unit connection network, the hybrid-order community detection can be designed by partitioning the micro-unit connection network into several micro-unit groups. From the micro-unit groups, the original node-wise community structure can be derived, i.e. nodes contained by one micro-unit group are assigned to the same node-wise community. A byproduct benefit of this approach is that the overlapping community structure can be discovered, as will be elaborated later.

The Micro-unit Modularity Model

In order to partition the micro-unit connection network into several disjoint micro-unit groups, denoted as $\mathcal{S}$, inspired by the classical modularity approach (Newman 2006), a micro-unit modularity model is designed. The goal is to find the intensively linked micro-unit groups. That is, for each group $s \in \mathcal{S}$, the *sum of micro-unit connection strength in group* s should be as large as possible compared with the *sum of expected micro-unit connection strength in group* s. The micro-unit modularity is designed for measuring how well the above goal is achieved,

$$Q \propto \sum_{s\in\mathcal{S}} \big((\text{sum of micro-unit connection strength in } s) - \gamma(\text{sum of expected micro-unit connection strength in } s)\big) \tag{4}$$

where γ is a resolution parameter and the second term involves a null model in the micro-unit connection network.

Similar to the classical modularity approach, we can construct a randomly rewired micro-unit connection network A' from the micro-unit connection network A by randomly rewiring the edges while keeping the same degree strength distribution of micro-units. Notice that, different from the null model in node-based networks, for each micro-unit, the adjacent micro-units can be categorized into two types, namely motif instances and edges. Therefore, for any two micro-units u_i and u_j, the expected micro-unit connection strength has the following four cases.

1. **Case I**: If both u_i and u_j are motif instances, their expected micro-unit connection strength A'_{ij} is defined as

$$A'_{ij} = \frac{d_i^{\mathbf{MM}} \cdot d_j^{\mathbf{MM}}}{2\mu^{\mathbf{MM}}} \tag{5}$$

where $d_i^{\mathbf{MM}}$ (*resp.* $d_j^{\mathbf{MM}}$) denotes the sum of connection strength of the adjacent motif instances of u_i (*resp.* u_j) and $\mu^{\mathbf{MM}} = \frac{1}{2}\sum_i d_i^{\mathbf{MM}}$.

2. **Case II**: If both u_i and u_j are edges, their expected micro-unit connection strength A'_{ij} is defined as

$$A'_{ij} = \frac{d_i^{\mathbf{EE}} \cdot d_j^{\mathbf{EE}}}{2\mu^{\mathbf{EE}}} \tag{6}$$

where $d_i^{\mathbf{EE}}$ (*resp.* $d_j^{\mathbf{EE}}$) denotes the sum of connection strength of the adjacent edges of u_i (*resp.* u_j) and $\mu^{\mathbf{EE}} = \frac{1}{2}\sum_i d_i^{\mathbf{EE}}$.

3. **Case III**: If u_i is a motif instance but u_j is an edge, their expected micro-unit connection strength A'_{ij} is defined as

$$A'_{ij} = \frac{d_i^{\mathbf{ME}} \cdot d_j^{\mathbf{EM}}}{2\mu^{\mathbf{ME}}} \tag{7}$$

where $d_i^{\mathbf{ME}}$ denotes the sum of connection strength of the adjacent edges of u_i, $d_j^{\mathbf{EM}}$ denotes the sum of connection strength of the adjacent motif instances of u_j, and $\mu^{\mathbf{ME}} = \frac{1}{2}\sum_i d_i^{\mathbf{ME}} = \frac{1}{2}\sum_j d_j^{\mathbf{EM}}$.

4. **Case IV**: The fourth case is similar to **Case III**.

Notice that, in the above four cases, the expected micro-unit connection strength A'_{ij} relies on the types of the micro-units u_i and u_j. In particular, it is not suitable to enumerate all adjacent micro-units when computing the expected micro-unit connection strength. For instance, in **Case I**, since both u_i and u_j are motif instances, the expected connection strength of u_i and u_j measures the expected strength of u_i connecting with u_j among all adjacent motif instances, which should not take into account the adjacent edges of u_i and u_j. Therefore, only $d_i^{\mathbf{MM}}$ is considered rather than $(d_i^{\mathbf{MM}} + d_i^{\mathbf{ME}})$. Similarly for the remaining cases.

According to the above discussion, Eq. (4) can be ex-

panded as

$$Q = \frac{1}{2\mu}\sum_{s\in\mathcal{S}}\sum_{u_i\in s}\sum_{u_j\in s}\Bigg(A_{ij} - \gamma\Bigg(\hbar(u_i\in\mathcal{M},u_j\in\mathcal{M})\frac{d_i^{\mathbf{MM}}\cdot d_j^{\mathbf{MM}}}{2\mu^{\mathbf{MM}}} + \hbar(u_i\in\mathcal{E},u_j\in\mathcal{E})\frac{d_i^{\mathbf{EE}}\cdot d_j^{\mathbf{EE}}}{2\mu^{\mathbf{EE}}} + \hbar(u_i\in\mathcal{M},u_j\in\mathcal{E})\frac{d_i^{\mathbf{ME}}\cdot d_j^{\mathbf{EM}}}{2\mu^{\mathbf{ME}}} + \hbar(u_i\in\mathcal{E},u_j\in\mathcal{M})\frac{d_i^{\mathbf{EM}}\cdot d_j^{\mathbf{ME}}}{2\mu^{\mathbf{ME}}}\Bigg)\Bigg) \quad (8)$$

where $\hbar(x,y)=1$ if both x and y are true, and 0 otherwise, and $\mu = \frac{1}{2}\sum_{i,j=1}^{|\{u\}|}A_{ij}$.

To further simplify the notation of the above equation, we introduce the micro-unit modularity matrix $B \in \mathbb{R}^{|\{u\}|\times|\{u\}|}$ with $|\{u\}|$ being the number of micro-units,

$$B_{ij} = A_{ij} - \gamma\Bigg(\hbar(u_i\in\mathcal{M},u_j\in\mathcal{M})\frac{d_i^{\mathbf{MM}}\cdot d_j^{\mathbf{MM}}}{2\mu^{\mathbf{MM}}} + \hbar(u_i\in\mathcal{E},u_j\in\mathcal{E})\frac{d_i^{\mathbf{EE}}\cdot d_j^{\mathbf{EE}}}{2\mu^{\mathbf{EE}}} + \hbar(u_i\in\mathcal{M},u_j\in\mathcal{E})\frac{d_i^{\mathbf{ME}}\cdot d_j^{\mathbf{EM}}}{2\mu^{\mathbf{ME}}} + \hbar(u_i\in\mathcal{E},u_j\in\mathcal{M})\frac{d_i^{\mathbf{EM}}\cdot d_j^{\mathbf{ME}}}{2\mu^{\mathbf{ME}}}\Bigg) \quad (9)$$

In addition, let $g_i, \forall i = 1,\ldots,|\{u\}|$ denote the group label of micro-unit u_i. Eq. (8) can be written as

$$Q = \frac{1}{2\mu}\sum_{i,j=1}^{|\{u\}|}B_{ij}\delta(g_i,g_j) \quad (10)$$

where $\delta(x,y)$ is the Kronecker delta, i.e. $\delta(x,y)=1$ if $x=y$ and 0 otherwise.

Similar to the classical modularity measure, finding intensively linked micro-unit groups requires finding the group labels of micro-units $g_i, \forall i = 1,\ldots,|\{u\}|$ by maximizing Eq. (10), which can be solved by the "generalized Louvain" approach (Jeub et al. 2011 2017). Notice that, the number of micro-unit groups k is automatically estimated by the "generalized Louvain" approach, i.e. the optimal k is set to be the number in which the maximum value of Q is achieved.

Algorithm Summary and Analysis

Based on the group labels of micro-units $g_i, \forall i = 1,\ldots,|\{u\}|$, the community structure of the nodes in the original network can be derived, which is denoted by $\{\mathcal{C}_1,\ldots,\mathcal{C}_k\}$, where k is the number of micro-unit groups, which is also the number of the predicted communities of the original network. In particular, all the nodes contained by the micro-units u_i belonging to the micro-unit group c, i.e. $g_i = c$, are assigned to community $\mathcal{C}_c$. That is,

$$\mathcal{C}_c = \{v_{i'}|v_{i'}\in\mathcal{V}^{u_i},\ \text{s.t. } g_i = c\}, \quad \forall c = 1,\ldots,k. \quad (11)$$

Algorithm 1 MuMod

Require: Network $\mathcal{G}=\{\mathcal{V},\mathcal{E}\}$, resolution parameter γ.
1: Discover motif by Definition 1.
2: Find the micro-unit set $\{u\}$ by Definition 5.
3: Construct the micro-unit connection network A by Definition 7.
4: Construct the micro-unit modularity matrix B by Eq. (9).
5: Obtain the group labels of micro-units $g_i, \forall i = 1,\ldots,|\{u\}|$ by maximizing Eq. (10) via the "generalized Louvain" approach.
6: Convert the group labels of micro-units $g_i, \forall i = 1,\ldots,|\{u\}|$ into the final communities $\{\mathcal{C}_1,\ldots,\mathcal{C}_k\}$ by Eq. (11).
Ensure: Communities $\{\mathcal{C}_1,\ldots,\mathcal{C}_k\}$.

For clarity, Algorithm 1 summarizes the main procedure of the proposed Micro-unit Modularity (MuMod) approach for hybrid-order community detection.

Compared with the existing lower-order community detection approaches, one advantage of MuMod is that the higher-order connectivity pattern is leveraged. On the other hand, compared with the existing higher-order community detection approaches, MuMod is able to address the hypergraph fragmentation issue by leveraging the lower-order connectivity pattern.

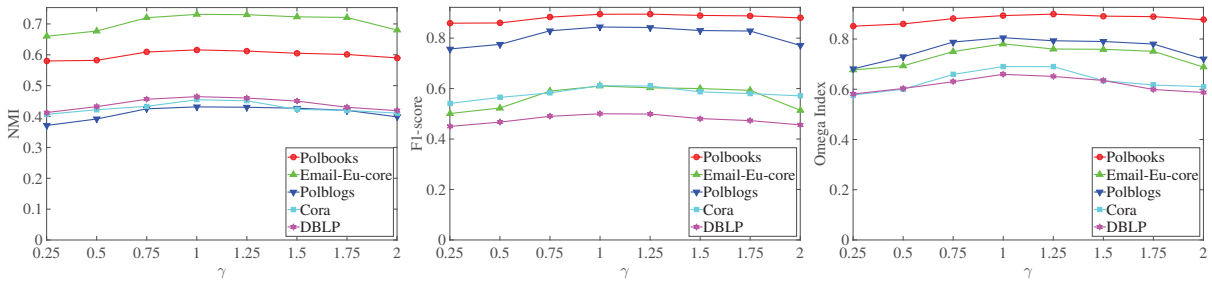
Another distinguishing merit is that, MuMod is able to capture the overlapping community structure. That is, when converting the group labels of micro-units $g_i, \forall i = 1,\ldots,|\{u\}|$ into the final communities $\{\mathcal{C}_1,\ldots,\mathcal{C}_k\}$ by Eq. (11), it is possible that one original singleton node $v_{i'}$ may be contained by more than one micro-units belonging to different micro-unit groups. That is, $\exists v_{i'}$, s.t. $v_{i'}\in\mathcal{V}^{u_{i_1}}, v_{i'}\in\mathcal{V}^{u_{i_2}}, u_{i_1}\neq u_{i_2}, g_{i_1}\neq g_{i_2}$. Different from the existing overlapping community detection approaches, MuMod does not encounter the ambiguity issue of overlapping communities, which is usually suffered by the existing approaches. For instance, the community membership strength matrix based approaches need to carefully tune some threshold to determine whether one node should be assigned to one community according to the community membership strength (Yang and Leskovec 2013; Huang, Wang, and Chao 2018b).

Experiments

Experimental Settings

Testing Datasets Five widely used real-world datasets are adopted as the testing datasets.

1. Polbooks[1]: A network of books about US politics consisting of 105 nodes and 441 edges, where nodes represent books and edges represent frequent co-purchasing of books by the same buyers. The 105 nodes are classified into 3 classes.



(a) NMI (b) F1-score (c) Omega Index

Figure 4: Parameter analysis on the effect of the resolution parameter γ on the performance of the proposed MuMod method.

2. Email-Eu-core[2]: An email network of communication between institution members consisting of 1005 nodes and 25571 edges, which is generated using email data from a large European research institution. The 1005 nodes are classified into 42 classes.
3. Polblogs[1]: A network of hyperlinks between weblogs on US politics consisting of 1490 nodes and 19090 edges, which is recorded in 2005. The 1490 nodes are classified into 2 classes.
4. Cora[3]: A citation network of scientific publications consisting of 2708 nodes and 5429 edges. The nodes are machine learning papers that can be classified into 7 classes.
5. DBLP[4]: A subset of the DBLP data. It consists of 11 premier research conferences in the fields of "DM&DB", "AI&ML" and "CV&PR" from 2001 to 2011, which are KDD, ICDE, ICDM, VLDB, SIGMOD, AAAI, IJCAI, ICML, NIPS, CVPR, ICCV and ECCV. It models co-author relationship between authors, where each node represents an author and each edge presents a co-author relationship between authors. Only those authors having no less than 5 papers published in these conferences from 2001 to 2011 were selected. There are overall 2554 nodes and 9963 edges. The nodes are classified into 3 overlapping communities according to the main fields of the published papers.

Comparison Methods Both of the lower-order community detection approaches and the higher-order community detection approaches are utilized as baselines.

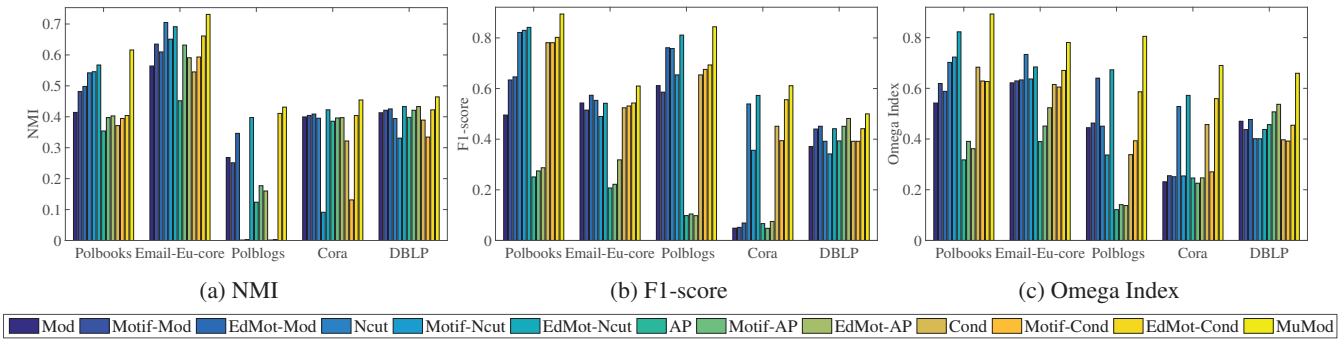
The following four lower-order approaches are adopted.

1. Modularity (Mod) (Newman 2006): It utilizes the "generalized Louvain" approach for maximizing the modularity measure.
2. Spectral clustering based on normalized cut (Ncut) (Shi and Malik 2000): It uses the spectral graph theory for minimizing the normalized cut measure.
3. Affinity propagation (AP) (Frey and Dueck 2007): It uses the classical affinity propagation algorithm for generating cluster labels.
4. Spectral clustering based on conductance (Cond) (Schaeffer 2007): It uses the spectral graph theory for minimizing the conductance measure.

The input of the above four lower-order approaches is the adjacency matrix of the original network. For each lower-order approach, a corresponding higher-order variant can be obtained and compared by taking as input the motif adjacency matrix rather than the original adjacency matrix. For instance, as described in (Benson, Gleich, and Leskovec 2016), by adopting the motif adjacency matrix as input to the classical conductance method, the Motif-Conductance (Motif-Cond) method can be obtained. Therefore, four motif-based higher-order approaches can be obtained, namely Motif-Mod, Motif-Ncut, Motif-AP and Motif-Cond. In addition, the recently developed edge enhancement method is adopted to construct the adjacency matrix, called EdMot (Li et al. 2019b), which is taken as input to the above four lower-order approaches, resulting in another type of higher-order approaches, namely EdMot-Mod, EdMot-Ncut, EdMot-AP and EdMot-Cond. For the above twelve comparison methods, the parameters are tuned as suggested by the original authors. In particular, for the higher-order approaches, like MuMod, the same 3-node triangle motif is adopted.

Evaluation Measures The three evaluation measures adopted in this paper include Normalized Mutual Information (NMI), F1-score and Omega Index. NMI is an information theory based evaluation measure for comparing the predicted communities and the ground-truth communities. Refer to (Strehl and Ghosh 2002) for details. F1-score is computed based on the mapping of the predicted communities and the ground-truth communities. Omega Index is suitable for measuring the quality of the overlapping community structure, which estimates the number of communities that each pair of nodes shares. For F1-score and Omega Index, refer to (Yang and Leskovec 2013) for details. For all the three measures, a higher value indicates better performance.

[1]http://www-personal.umich.edu/~mejn/netdata/
[2]http://snap.stanford.edu/data/
[3]http://linqs.cs.umd.edu/projects/projects/lbc/
[4]http://dblp.uni-trier.de/



(a) NMI (b) F1-score (c) Omega Index

Figure 5: Comparison results with the twelve existing methods on the five real-world networks.

Parameter Analysis

In this section, parameter analysis will be conducted to show how the resolution parameter γ would affect the performance of MuMod. To this end, on each of the five testing networks, we run MuMod by setting $\gamma \in \{0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2\}$ and report the performance in terms of NMI, F1-score and Omega Index. The results are shown in Figure 4. From the figure, we can see that in general MuMod is relatively stable in the testing range of γ. In particular, when setting γ as 1 and 1.25, the best results can be obtained, which coincides with the studies in the previous work (Newman 2006; Mucha et al. 2010). Following (Mucha et al. 2010), we will set the resolution parameter γ as 1 in the following experiments.

Comparison Results

In this section, comparison experiments will be conducted to compare MuMod with the twelve methods, including four lower-order approaches and their eight higher-order variants. Comparison results in terms of NMI, F1-score and Omega Index are reported in Figure 5. From the figure, we can see that in general, compared with the lower-order approaches, the higher-order variants obtain much better results. In general, more than 10% improvements in terms of NMI have been achieved by the eight higher-order variants over the four lower-order approaches. This has confirmed the benefit of leveraging the higher-order connectivity pattern in community detection. However, in some cases, the higher-order approaches are not as good as the lower-order approaches. For instance, in the Cora network, the two motif-based higher-order approaches, namely Motif-Ncut and Motif-Cond, obtain the NMI values that are much smaller than their lower-order counterparts, namely Ncut and Cond. This is mainly caused by the hypergraph fragmentation issue (Li et al. 2019b), which results in the community structure consisting of a large number of over-segmented communities. Although the recent edge-enhancement based approaches, namely EdMot based variants, can reduce the impact caused by the hypergraph fragmentation issue and improve the performance in most cases compared with both of the original lower-order methods and the motif-based higher-order methods. However, adding additional edges would destroy the lower-order connectivity pattern, which makes the edge enhancement based approaches fail to generate better results in some networks, e.g. Email-Eu-core and Polblogs.

As a comparison, MuMod generates the best results in terms of NMI, F1-score and Omega Index on all the testing datasets. In particular, at least 15% improvement in terms of NMI has been obtained. Similar quantity analysis can be made in terms of F1-score and Omega Index. The main reason lies in that MuMod makes full use of both of the lower-order connectivity pattern and the higher-order connectivity pattern by means of the micro-unit connection. Therefore, it can effectively address the hypergraph fragmentation issue without destroying the lower-order connectivity pattern. In addition, MuMod has the capability of capturing the overlapping community structure, which can be especially reflected from the values of Omega Index on DBLP.

Conclusion

In this paper, we have defined a new community detection problem called hybrid-order community detection and proposed a new Micro-unit Modularity (MuMod) approach. Different from the existing community detection approaches, both of the lower-order connectivity pattern and the higher-order connectivity pattern are utilized by means of constructing a micro-unit connection network. A micro-unit modularity model is designed for generating the micro-unit groups, based on which the overlapping community structure of the original network is derived. Experiments have confirmed the effectiveness of the proposed method.

Acknowledgments

This project was supported by NSFC (61672548, U1611461).

References

Arenas, A.; Fernández, A.; Fortunato, S.; and Gómez, S. 2008. Motif-based communities in complex networks. *Journal of Physics A: Mathematical and Theoretical* 41(22):224001.

Benson, A. R.; Gleich, D. F.; and Leskovec, J. 2016. Higher-order organization of complex networks. *Science* 353(6295):163–166.

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of community hierarchies in large networks. *J. Stat. Mech.* 2008(10):P10008.

Chakraborty, T.; Srinivasan, S.; Ganguly, N.; Mukherjee, A.; and Bhowmick, S. 2014. On the permanence of vertices in network communities. In *KDD*, 1396–1405.

Frey, B. J., and Dueck, D. 2007. Clustering by passing messages between data points. *Science* 315:972–976.

Ganji, M.; Bailey, J.; and Stuckey, P. J. 2018. Lagrangian constrained community detection. In *AAAI*, 2983–2990.

He, L.; Lu, C.-T.; Ma, J.; Cao, J.; Shen, L.; and Yu, P. S. 2016. Joint community and structural hole spanner detection via harmonic modularity. In *KDD*, 875–884.

He, D.; Feng, Z.; Jin, D.; Wang, X.; and Zhang, W. 2017. Joint identification of network communities and semantics via integrative modeling of network topologies and node contents. In *AAAI*, 116–124.

He, D.; You, X.; Feng, Z.; Jin, D.; Yang, X.; and Zhang, W. 2018. A network-specific markov random field approach to community detection. In *AAAI*, 306–313.

Huang, L.; Wang, C.-D.; and Chao, H.-Y. 2018a. A harmonic motif modularity approach for multi-layer network community detection. In *ICDM*, 1043–1048.

Huang, L.; Wang, C.-D.; and Chao, H.-Y. 2018b. Overlapping community detection in multi-view brain network. In *BIBM*, 655–658.

Huang, L.; Wang, C.-D.; and Chao, H.-Y. 2019. Higher-order multi-layer community detection. In *AAAI*, 9945–9946.

Jeub, L. G. S.; Bazzi, M.; Jutla, I. S.; and Mucha, P. J. (2011-2017). A generalized Louvain method for community detection implemented in MATLAB. http://netwiki.amath.unc.edu/GenLouvain.

Jin, D.; Wang, X.; He, R.; He, D.; Dang, J.; and Zhang, W. 2018. Robust detection of link communities in large social networks by exploiting link semantics. In *AAAI*, 314–321.

Jin, D.; You, X.; Li, W.; He, D.; Cui, P.; Fogelman-Soulié, F.; and Chakraborty, T. 2019. Incorporating network embedding into markov random field for better community detection. In *AAAI*, 160–167.

Laishram, R.; Wendt, J. D.; and Soundarajan, S. 2019. Crawling the community structure of multiplex networks. In *AAAI*, 168–175.

Li, J.-H.; Wang, C.-D.; Li, P.-Z.; and Lai, J.-H. 2018a. Discriminative metric learning for multi-view graph partitioning. *Pattern Recognition* 75:199–213.

Li, P.-Z.; Huang, L.; Wang, C.-D.; Huang, D.; and Lai, J.-H. 2018b. Community detection using attribute homogenous motif. *IEEE ACCESS* 6:47707–47716.

Li, P.-Z.; Cai, Y.-X.; Wang, C.-D.; Liang, M.-J.; and Zheng, Y.-Q. 2019a. Higher-order brain network analysis for auditory disease. *Neural Processing Letters* 49:879–897.

Li, P.-Z.; Huang, L.; Wang, C.-D.; and Lai, J.-H. 2019b. EdMot: An edge enhancement approach for motif-aware community detection. In *KDD*, 479–487.

Lin, W.; Xiao, X.; Xie, X.; and Li, X. 2017. Network motif discovery: A GPU approach. *IEEE TKDE* 29(3):513–528.

Milo, R.; Shen-Orr, S.; Itzkovitz, S.; Kashtan, N.; Chklovskii, D.; and Alon, U. 2002. Network motifs: simple building blocks of complex networks. *Science* 298(5594):824–827.

Milo, R.; Itzkovitz, S.; Kashtan, N.; Levitt, R.; Shen-Orr, S.; Ayzenshtat, I.; Sheffer, M.; and Alon, U. 2004. Superfamilies of evolved and designed networks. *Science* 303(5663):1538–1542.

Mucha, P. J.; Richardson, T.; Macon, K.; Porter, M. A.; and Onnela, J. 2010. Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. *Science* 328(5980):876–878.

Newman, M. E. 2006. Modularity and community structure in networks. *PNAS* 103(23):8577–8582.

Schaeffer, S. E. 2007. Graph clustering. *Computer Science Review* 1(1):27–64.

Shao, J.; Han, Z.; Yang, Q.; and Zhou, T. 2015. Community detection based on distance dynamics. In *KDD*, 1075–1084.

Shi, J., and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE TPAMI* 22(8):888–905.

Strehl, A., and Ghosh, J. 2002. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *JMLR* 3:583–617.

Sun, B.-J.; Shen, H.; Gao, J.; Ouyang, W.; and Cheng, X. 2018. Towards efficient detection of overlapping communities in massive networks. In *AAAI*, 418–425.

Tsourakakis, C. E.; Pachocki, J.; and Mitzenmacher, M. 2017. Scalable motif-aware graph clustering. In *WWW*, 1451–1460.

Wang, C., and Zhu, J. 2019. Forbidden nodes aware community search. In *AAAI*, 758–765.

Wang, C.-D.; Lai, J.-H.; and Yu, P. S. 2013. Dynamic community detection in weighted graph streams. In *SDM*, 151–161.

Wang, C.-D.; Lai, J.-H.; and Yu, P. S. 2014. NEIWalk: Community discovery in dynamic content-based networks. *IEEE TKDE* 26(7):1734–1748.

Yang, J., and Leskovec, J. 2013. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *WSDM*, 587–596.

Yin, H.; Benson, A. R.; Leskovec, J.; and Gleich, D. F. 2017. Local higher-order graph clustering. In *KDD*, 555–564.

Zhang, H.; Wang, C.-D.; Lai, J.-H.; and Yu, P. S. 2019. Community detection using multilayer edge mixture model. *Knowl. Inf. Syst.* 60(2):757–779.

Zhao, P. 2015. gSparsify: Graph motif based sparsification for graph clustering. In *CIKM*, 373–382.

Zhou, D.; Zhang, S.; Yildirim, M. Y.; Alcorn, S.; Tong, H.; Davulcu, H.; and He, J. 2017. A local algorithm for structure-preserving graph cut. In *KDD*, 655–664.